基于 Sas 的时间序列缺失值处理方法比较

兰 妥[1],江 弋[1],刘光生[2]
(1. 厦门大学 信息学院 计算机系,福建 厦门 361005;
2. 兰州大学 资源环境学院,甘肃 兰州 730000)

摘 要:对于时间序列挖掘过程中的缺失值处理,目前有许多方法。在处理数据变量成一定的相关的数据集时,回归模型不失为较好的插补方法。利用均值插补、一元线性回归、多元线性回归、迭代回归方法对水文时间序列数据集的缺失数据进行处理,比较不同的皮氏相关系数下各方法的优劣及适用性。文中研究表明当数据集中存在与缺值变量相关度较大的变量时,一元线性回归的插补简单直观,且有较高的精度,结果接近真实;当数据集中不存在与缺值变量显著相关的自变量时,一元线性回归的结果变差,多元线性回归与多元迭代回归具有较好的结果,但多元迭代回归迭代次数难以确定,插补代价较大,多元线性回归为最佳选择;当缺值变量与其他自变量相关系数均较小时,回归插补的结果不理想,此时可考虑其他插补方法。

关键词:时间序列;皮氏相关系数;回归模型;迭代回归模型
中图分类号:TP311 **文献标识码**:A **文章编号**:1673 - 629X(2008)10 - 0043 - 03

Comparison of Methods on Time Series' Missing Value Based on Sas

LAN Tuo[1], JIANG Yi[1], LIU Guang-sheng[2]
(1. Dept. of Computer Science, Info. Sch., Xiamen University, Xiamen 361005, China;
2. School of Resources and Environment, Lanzhou University, Lanzhou 730000, China)

Abstract: There are many methods for dealing with missing value on time series data. When the variables of the data are correlative, the regression model is better than other methods. Handles missing value of hydrological by using mean interpolation, single linear regression, multiple linear regression and interative regression method. Shows that when the data set exists the variable which relates with given variable closely, the single linear regression is better than other methods. If the data set doesn't, multiple linear regression is best. If the pearson correlation between the given variable and other variables is small, may consider other interpolation method.

Key words: time series; pearson correlation; one - stepregression model; iterative regression model

0 引 言

时间序列是按照时间顺序取得的一系列观测值,它是参数离散的随机过程[1]。在实际中,遇到的许多统计资料,如某地的月降雨量、某年高原环境观测系统数据、某工厂装船货物数量的月度序列、某交通口的日车流量,都是时间序列。

但在许多实际情况中,往往存在着缺失值的问题,为此,时间序列分析者们进行了大量的工作,提出了许多处理缺失数据的方法[2]。常见的插补模型如随机抽取替代模型、均值替代模型、最近临域替代模型、多重插补、基于 EM 算法的替代模型和回归模型等[3,4]。

在处理数据变量成一定的相关的数据集时,回归模型的效果不失比其他的方法好。回归模型包括一元线性回归、多元线性回归迭代回归。文中利用这均值插补和这三种方法对时间序列缺失数据插补,比较在不同皮氏相关系数下的插补结果。

1 皮氏相关系数

通过计算两个属性 A 和 B 之间的相关系数,可以估计这两个属性的相关度

$$r_{A,B} = \frac{\sum_{i=1}^{N}(a_i - \bar{A})(b_j - \bar{B})}{N\sigma_A\sigma_B}$$

其中 N 是元组个数,a_i 和 b_j 分别是元组 i 中 A 和 B 的值,$\bar{A}$ 和 $\bar{B}$ 分别是 A 和 B 的均值,σ_A 和 σ_B 分别是 A 和

收稿日期:2008 - 01 - 16
基金项目:福建省自然科学基金资助项目(A031008)
作者简介:兰 妥(1984 -),女(畲族),福建古田人,硕士研究生,研究方向为数据库应用、时间序列数据挖掘;江 弋,副教授,硕士生导师,研究方向为数据库技术与应用、数据挖掘、电子商务、多媒体技术及应用、嵌入式系统。

B 的标准差，$-1 \le r_{A,B} \le +1$，如果 $r_{A,B}$ 大于0，则 A 和 B 正相关，否则负相关[5]。

2 回归模型

鉴于时间序列的特性，文中采取回归模型对缺失数据插补，分别利用一元线性回归、多元线性回归、迭代回归进行插补。

2.1 一元线性回归

通过构建因变量与某一相关自变量之间的一元回归模型，给出合理的回归方程，进而估算缺失数据，假设数据集中的第 i 个样本中第 j 个元素是缺失数据，即

$$x_i = (x_{i1}, x_{i2}, \cdots, x_{ij-1}, x_{ij+1}, \cdots, x_{im})$$

如能够求得不完全变量 v_j 同某一相关变量的回归方程式：

$$v_j = u_0 + \mu v_i$$

则可以简单高效地得出缺失数据的取值，一元线性回归的任务是求出最佳线性拟合，通常采用最小二乘法求得。

2.2 多元回归模型

1992年，Albrecht 研究了多元回归模型在处理缺失数据中的应用[6]，对变量明显的相关的数据集处理时，其效果通常比其他的统计方法更好、更直接。它通过回归分析构建因变量与自变量之间的回归模型，给出合理的回归方程，进而估算缺失数据。假设数据集中的第 i 个样本中第 j 个元素是缺失数据，即

$$x_i = (x_{i1}, x_{i2}, \cdots, x_{ij-1}, x_{ij+1}, \cdots, x_{im})$$

如能够求得不完全变量 v_j 同其他变量的回归方程式

$$v_j = u_0 + \sum_{l:\, l=1,\ l \ne j}^{m} \mu_l v_l + \varepsilon$$

就很容易预测出缺失数据的取值。回归模型预测缺失数据，必须建立一个合理的回归方程式，或者是建立一个变量变换后的回归方程式(将非线性模型转化成线性模型处理)。

2.3 迭代回归模型

回归模型通常只需建立一次回归方程式，但是为了得到缺失数据更精确的估计值，1968年，Jackson 提出用迭代回归模型来处理缺失数据[7]。迭代回归模型的主要思想就是基于均值替代模型的结果建立回归模型来估算缺失值，反复迭代并估算出缺失值，直到前后两次的估计值改变量小于事先给定的阈值为止。通常迭代回归估算缺失值比一步回归估算缺失值更准确，但其算法比一步回归估算复杂，其思路可以归纳如下(但是过多地迭代会导致大的误差，而且计算效率降低)：

(1)给定数据集中的缺失值，基于均值替代生成完整的数据集，即

$x_{ij} = \bar{x}_j$，其中 $\bar{x}_j = \dfrac{\sum_{i=1}^{n} x_{ij}}{n}$，这里 n 为完整样本个数；

(2)生成完整的数据集，建立合理的回归模型并进行缺失部分的数据估计并替代；

(3)根据回归替代生成的数据集和回归式子反复进行替代，直到前后两次回归估算值相差小于某个值时，即 $\sum | x_{ij}^{(i)} - x_{ij}^{(i+1)} | < \delta$，$\delta$ 为事先给定的阈值，回归结束。

3 实验过程

选取2004年高原 ENVIS 环境观测系统数据，共1440个观测样本，24个变量，对完整数据集构造不同变量的随机缺失数据集，分别采用均值插补，一元线性回归、多元回归插补方法进行插补。对插补结果利用评分函数 e 的分布和 E 的大小来衡量。

$$e = \frac{| x_{ij} - x'_{ij} |}{x_{ij}},\quad E = \frac{\sum_{x_{ij} \in M} | x_{ij} - x'_{ij} |}{\sum_{x_{ij} \in M} x_{ij}}$$

其中 x_{ij} 表示真实值，x'_{ij} 表示运用插补生成的替代值，M 为所有缺失值构成的集合。

利用 sas 的 corr 过程，计算缺值变量气温与其他变量的相关系数[8]。其中向上长波辐射与气温相关系数最大，皮氏相关系数 $r_{A,B} = 0.81970$。分别采用均值插补，一元回归方法(选取向上长波辐射作为自变量构建一元回归模型)，多元回归方法(选取其余的23个指标作为自变量构建多元回归模型)，迭代回归方法进行插补，利用评分函数 e 和 E 对比插补结果如表1和表2所示。

表1 e 值分布(单位：个)

插补模型 e	e	0-0.05	0.05-0.1	0.1-0.2	0.2-0.3	0.3-high
均值替补		8	11	15	12	14
一元回归		18	15	18	4	5
迭代回归(二步)		17	16	18	4	5
多元回归		19	23	15	6	7

表2 E 值

插补模型	均值替补	一元回归	迭代回归	多元回归
E	0.2071090377	0.105797437	0.1057786833	0.103959807

利用 sas 的 corr 过程，计算缺值变量风速与其他变量的相关系数[8]。其中向上短波辐射变量与其相关度最大，$r_{A,B} = 0.63415$。分别采用均值插补，一元回归方法(选取向上短波辐射作为自变量构建一元回归

模型),多元回归方法(选取其余的23个指标作为自变量构建多元回归模型),迭代回归方法进行插补,利用评分函数 e 和 E 对比插补结果,如表3和表4所示。

表3　e 值分布(单位:个)

插补模型 e	e	0 - 0.05	0.05 - 0.1	0.1 - 0.2	0.2 - 0.3	0.3 - high
均值替补		4	9	13	5	29
一元回归		4	6	15	11	24
迭代回归(二步)		8	12	24	7	9
多元回归		8	10	21	9	12

表4　E 值

插补模型	均值替补	一元回归	迭代回归	多元回归
E	0.3697505998	0.3142652147	0.189320354	0.190217811

利用 sas 的 corr 过程,计算缺值变量湿度与其他变量的相关系数[8],其中气压与湿度相关度最大,$r_{A,B} = 0.45442$。分别采用均值插补,一元回归方法(选取气压作为自变量构建一元回归模型),多元回归方法(选取其余的23个指标作为自变量构建多元回归模型),迭代回归方法进行插补,利用评分函数 e 和 E 对比插补结果,如表5和表6所示。

表5　e 值分布(单位:个)

插补模型 e	e	0 - 0.05	0.05 - 0.1	0.1 - 0.2	0.2 - 0.3	0.3 - high
均值替补		4	5	5	10	36
一元回归		2	1	10	9	38
迭代回归(二步)		2	3	14	9	32
多元回归		2	3	9	6	32

表6　E 值

插补模型	均值替补	一元回归	迭代回归	多元回归
E	0.4538440755	0.419867311	0.3361965374	0.3362214112

4　结束语

在变量成一定相关性时间序列数据集中,回归模型是缺失值处理中较好的模型,实验表明回归模型比传统的均值插补等效果好的多,但不同的回归方法对不同的数据集优劣性与适用性不同。

本实验研究证明:

若数据集中存在与缺值变量显著相关的自变量,皮氏相关系数大于0.8时,一元线性回归插补方法简单高效,且插补结果较好,此时多元回归与迭代回归插补结果没有显著优越性,且占用的内存和时间较大,因此不是最优选择。

若数据集中不存在与缺值变量显著相关的自变量,皮氏相关系数均0.5～0.8之间,一元回归效果明显不如多元回归与迭代回归,此时多元回归与迭代回归插补结果相当,考虑多元回归的时空代价性相对迭代回归小,此时多元回归可视为最优选择。

若缺值变量与其他自变量相关系数均较小,皮氏相关系数均小于0.5时,一元回归、迭代回归、多元回归的效果与均值插补的效果相当,都不理想,此时应该考虑其他插补方法,例如多重插补、基于聚类的插补等等。

参考文献:

[1] Brockwell P J, Davis R A. 时间序列的理论与方法[M]. 北京:高等教育出版社,2001.

[2] Little R J A, Rubin D B. 缺失数据统计与分析[M]. 孙　山译. 北京:中国统计出版社,2004.

[3] 金勇进. 缺失数据的插补调整[J]. 数理统计与管理,2001(5):47 - 53.

[4] 刘　鹏,雷　蕾,张雪凤. 缺失数据处理方法的比较研究[J]. 计算机科学,2004(9):155 - 156.

[5] Han Jiawei, Kamber M. 数据挖掘:概念与技术[M]. 第2版. 范　明等译. 北京:机械工业出版社,2007.

[6] Albrecht G H. Multivariate morphometrics with missing data: techniques for canonical variates and generalized distances[J]. Am J phys. Anthropol,1992,14(S14):42 - 48.

[7] Jackson E C. Missing values in linear multiple discriminant analysis[J]. Biometrics,1968,23:835 - 844.

[8] 朱世武. SAS 编程技术与金融数据处理[M]. 北京:清华大学出版社,2003.

(上接第42页)

方法[J]. 量子电子学报,2007,24(1):7 - 12.

[2] 刘凯峰,张德祥. 基于小波变换区域方差的遥感图像融合新算法[J]. 计算机技术与发展,2007,17(5):177 - 179.

[3] Rioul O, Flandrin P. Time - Scale Energy Distributions: A General Class Extending Wavelet Transforms [J]. IEEE Transactions on Signal Processing,1992,40(7):1746 - 1757.

[4] Cohen L. Time - Frequency Distributions - A Review[J]. IEEE Proceedings,1989,77(7):941 - 981.

[5] 罗利春. 谱相关的原理、功能与截面谱表示[J]. 物理学报,2002,51(10):2177 - 2182.

[6] 黄春琳,姜文利,周一宇. 低截获概率雷达信号的循环谱相关函数检测方法分析[J]. 国防科技大学学报,2001,23(4):102 - 106.

[7] 王李军,熊　刚,赵惠昌,等. 基于小波的广义时频分布及其与 Cohen 类的关系[J]. 电子与信息学报,2005,27(12):1927 - 1932.

[8] 张仔兵,李立萍,肖先赐. MPSK 信号的循环谱检测及码元速率估计[J]. 系统工程与电子技术,2005,27(5):803 - 806.